REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
14-08-2014	FINAL REPORT	Nov 2013 - Jun 2014

4. TITLE AND SUBTITLE

"Smart De-Identification of Social Media Data: Final Report"

5a. CONTRACT NUMBER

N00014-14-P-1089

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

1

6. AUTHOR(S)

Carley, Kathleen M.
Carley, L. Richard
Storrick, Jonathan

5d. PROJECT NUMBER

5e. TASK NUMBER

1

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Carley Technologies Inc.
128 N. Craig St., Suite 210
Pittsburgh, PA 15213 USA

8. PERFORMING ORGANIZATION REPORT NUMBER

Smart_DeIdent_PO_1089_v5_rpt6

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Office of Naval Research
Attn: Rebecca Goolsby, ONR Code 341
One Liberty Center
875 N. Randolph Street
Arlington, VA 22203-1995

10. SPONSOR/MONITOR'S ACRONYM(S)

ONR

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

Distribution Statement A. Approved for Public Release. Distribution is unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Final Technical and Financial Report on the development of a prototype de-identification tool for Twitter. The objective of this work was to develop an understanding of the full range of requirements for a de-identification tool to remove personally identifiable information from social media data, both the structured and the un-structured components, to design such a tool and develop a prototype. The resulting de-identifier is scalable, supports multiple levels of de-identification, and enables smart-de-identification (and so retention of object class information) in a fashion that is under control of the user, supports over-time and cross-data comparison, and meets legal requirements.

15. SUBJECT TERMS

De-Identification, Personally Identifying Information

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Larry Richard (Rick) Carley
U	U	U	UU	6	19b. TELEPHONE NUMBER *(Include area code)* 412-953-8818

Reset

Project Title: "Smart De-Identification of Social Media Data"

Report Title: Final Technical and Financial Report

Navy PO #: N00014-14-P-1089

Award Period: 21 Nov 2013 – 13 Jun 2014

Award Amount: $149,301.51

Report Date: 13 Jun 2014

Submitted to: Dr. Rebecca Goolsby, Code 341, ONR

PI: Dr. Kathleen M. Carley
Organization: Carley Technologies Inc.
1924 Glen Mitchell Road
Sewickley PA 15143 USA

e-mail: kathleen.carley@carleytech.com
Phone: 412-741-2002

Contracts Contact: Dr. L. R. Carley
Organization: Carley Technologies Inc.
1924 Glen Mitchell Road
Sewickley PA 15143 USA

e-mail: rick.carley@carleytech.com
Phone: 412-953-8818

Smart De-Identification of Social Media Data

Final Technical and Financial Report, 13 Jun 2014

Introduction

The objective was to develop an understanding of the full range of requirements for a de-identification tool to remove personally identifiable information from social media data, both the structured and the un-structured component, to design such a tool and develop a prototype for use with a social media. The de-identifier we developed is scalable, supports multiple levels of de-identification, and enables smart-de-identification (and so retention of object class information) in a fashion that is under control of the user, supports over-time and cross-data comparison, and meets legal requirements. The central challenge was to support de-identification while retaining exportability to standard statistical and network analytic packages and without destroying the structure of the data and so reducing the value of the data to address core questions. Central to the approach is the use characteristic matching, extensible database storage, parallelized match-replacement algorithms. Secondary information de-identification considers physical and social network characteristics in addition to personal contact information such as address, phone, and email-ids. The core focus of this effort was de-identification within the context of Twitter with applications in the areas of disaster response and crisis management.

Technical Progress

The outcome of this effort was the design and demonstration of a social-network de-identification system that, when applied to social media data, particularly Twitter, enables the social media analytic systems that use it to meet the requirements for a personal information assessment and so be compliant with DoD 5400.11-R and DoD 5400.11 yet to still support social network analytics and visualization. As such, the prototype system removes PII in compliance with Section 208 of Public Law 107-347; and OMB M-03-22.

Social media data presents a somewhat unique set of challenges in the area of de-identification. To understand these challenges several key concepts need to be defined:

Direct identifiers: These include those identifiers that unequivocally and on their own support knowing who is the person in question. Direct identifiers include the name, address and in some cases email handle and webpage of the individual.

Indirect (Quasi) identifiers: These include those identifiers that can collectively be used to re-identify an individual even if the direct information has been obfuscated or masked. This includes things like dates when the individual did something, locations where the individual was, and socio-demographic information such as income or birth date.

Anonimization: Anonimization refers to the permanent removal of personal identifiers from data such that once the personally identifiable information is removed from the data they can never be re-associated with the underlying individual.

Masking: Masking refers to the process of manipulating direct identifiers, usually by substituting a random id, using either a reversible or irreversible aliasing approach. When data is masked indirect or quasi identifiers are still present and could, in theory, be used for re-identification.

The key challenge is "big data." Consider these statistics from http://www.dazeinfo.com/2013/01/10/social-media-statistics-2013-facts-figures-facebook-twitter/ :

- 175 million tweets per day in 2012
- Twitter has more than half a billion registered profiles
- Twitter has 140 million profiles in the US

Any de-identification technology must be able to support the de-identification of the volume of data being produced, and with the recognition that the volume could increase. To meet this challenge we developed a system that could be ported to make use of cloud technologies.

The second key challenge, is near real time processing requirement. For de-identified social media data to be of value to the DoD, the de-identification must be done on a message by message basis, in as close to real-time as possible. Quasi-identifiers (e.g., dates and locations) emerge over time and do not come packaged with the direct identifiers (e.g. name and address). This is a fundamentally different problem than the de-identification of a database for sharing where the data can be frozen and de-identification proofs be done on that specific data. To meet this challenge we demonstrated a staged approach that focuses on masking data. We also explored the quantity of quasi-identifiable information that is present to develop a strategy for the statistical obfuscation of such information.

The third key challenge is that in social media, particularly in "message traffic" the actors are not all human individuals. For example, in Twitter, the tweeter might be a human, an organization, or a bot. These organizations can include news-organizations. While US laws inhibit certain military units from collection PII on US citizens, they are less restrictive with respect to companies. Further, knowing whether the information is coming from an individual or an organization, particularly a news organization, is critical for many missions. To meet this challenge we propose a context sensitive smart de-identification approach that can be used to de-identify all actors, or to retain class identification, or to retain all identifying information for just those actors in specific classes. Thus, we developed a system in which PII on human individuals could be removed but the names of actors that are news organizations, such as BBC Breaking news, could be retained.

The fourth key challenge is that for the de-identified data to be useful, the network structure of the data must be preserved at least at the "node-class" level. That is, while de-identification should prevent being able to specifically identify that Joe Black sent a particular tweet it should not remove the fact that a particular tweet was from a particular country or from a person versus a news-organization or bot. In a similar vein the temporal network structure needs to be preserved, that is the masks used between data sets and at specific time periods need to be consistent. To meet this challenge we tested the de-identified data with existing statistical and network tools to determine which categories of analytic and visualization functions are possible with different levels of de-identification and to use a context-specific smart de-identification scheme that supports the retention of class, rather than personal, information.

The fifth key challenge is the nature of the social media technology itself. These technologies are not stable and do change their APIs for extracting information from them, and even what information they extract from and save on the users. Further, each social media technology has somewhat different usage agreements. As such: a) there is no common API for extracting information across all social media; and b) what PII information is available can change over time. Different de-identification technologies are thus needed for different social media. To address this challenge we (a) gathered general requirements for a de-identification system from the operational community, (b) developed de-identification technology appropriate for Twitter; (c) demonstrated the use of this de-identifier at two different sponsor events, (d) explored what needs to be done for a similar technology for Facebook; and (e) developed a plan for which parts of the de-identification technology are transferable across diverse social media.

Thus, the outcome of this effort was a (a) a set of system requirements, and (b) a working tool de-identification of PII given twitter data. This working tools, should be thought of as a prototype smart-deidentification context-based technology for Twitter that will support multiple levels of de-identification. The three basic levels of de-identification supported are:

1) General masking – in this case names, addresses, webpages, and email ids are removed and a single alias for the individual substituted.
2) Class-based masking – in this case the individual and so the PII is categorized as to whether the actor is an individual, an organization, or a known news organization. The actor's PII is masked as in the general case but the aliases are chosen from class-categories – person, organization, news-organization and unknown.
3) Selective class-based masking – in this case, for those classes where it is ok to retain PII information such as for news agencies, the PII of the actor would not be masked; but, the PII for critical classes such as human individuals would be de-identified.

The developed social-network de-identification technology is extensible, flexible and scalable to the "big data" available vis social media, utilizes "human-friendly" versus mechanistic aliases, utilizes alias thesauri to track those classes of actors that can be selectively not masked, supports user selection of de-identification level, and supports alias-to-real mapping that enables the development of social networks and temporal social networks. The system is capable of operating in a distributed architecture to support big data processing. A secondary feature is the ability to set the system to provide the data in an anonymized fashion or just de-identified.

A secondary objective was the exploration of the utilization of quasi identifiers in social media in order to assess the feasibility of removal of such indirect identifiers. This is critical for it is only by knowing the prevalence and feasibility of removing quasi identifiers that fully operational tactics, techniques and procedures (TTPs) for handling social media in a way that supports mission needs can be developed. In other words TTPs that can be actually used without potentially dire unintended consequences cannot be developed without having this underlying knowledge. Based on our initial work, quasi identifiers are quite prevalent, but can be successfully removed. These include references to age, gender, and the incorporation of images. The problem, however, is that these quasi-identifiers might not refer to the message sender, and determining whom the identifiers apply is extremely hard. Basic research is

required to develop algorithms on whether it is possible, or the extent to which it is possible to identify which key-entity the identifiers refer to, and then whether it is necessary to do any de-identification. It is also important to note that quasi-identifiers are often used for determining factors such as race, age, gender and country-of-origin. However, such, "identification" is considered a largely unsolved problem for social media; when objective fields with that information do not exist in the meta-data. What this means is that to de-identify all and only US citizens you first need to identify the entity as a US citizen and then de-identify it. Basic research is needed on how to do key-entity identification sufficient for selective de-identification.

A third objective was to assess the scalability of the underlying technologies using the proposed prototype social-network de-identification system at various levels of scale. We were successful by developing a prototype that can be scaled across a Cloud server environment. Basic timing information on processing speed was collected for Twitter data. For many twitter providers, such as TweetTracker, de-identification adds negligible delay. It often takes longer to collect the data than to de-identify it.

Recommended Future Research Directions

The De-Identification Tool delivered under this research effort represents an initial exploration of the general problem of de-identification of social media sites. We recommend that additional research will be needed to handle the de-identification of a wide range of social media web sites, to develop pre-identification algorithms, and to do selective masking of quasi-identifiers.

Quasi or secondary identifiers were found to pose potentially severe problems for usable de-identification. At a surface level, it is easy to mask this information. When this is done indiscriminately all such information becomes effectively unavailable. If there are mission objectives that require smart de-identification then general quasi-identifier masking will obfuscate required data for meeting mission objectives. Examples of mission objectives that require smart de-identification are ones that require only US citizens to be de-identified but to leave non US citizens identified, or that require identification of discussions of general activity such as human trafficking. General obfuscation, prior to entity identification, will make it impossible to check for references to children, girls, boys, in human-trafficking posts. It will reduce the ability to identify which agents are actually not US citizens. Basic research is thus needed on the relation of key-entity socio-demographic classification by quasi-identifiers and the impact of that de-identification on classification.

The goal of this research effort was to demonstrate that de-identification is possible. We demonstrated that it was possible and that high level de-identification results in data that is still analyzable and that can be re-identified. We recommend that additional research on carrying out de-identification in the Cloud using massively parallel techniques would be needed in order to handle full scale data such as the Twitter Firehose, Youtube meta-data, or public-facebook. Future research should be done on what are the implications of not first creating a common identify for each agent across social-media before de-identification.

Future research needs to be conducted on the impact of the level of de-identification on what statistics and analyses can be run. That is, the biases created by de-identification need to

be assessed. Several classes of metrics were identified as being impacted by the style and level of de-identification: these include, but are not limited to, social network metrics, socio-demographic based atmospherics, organizational and news analytics. Basic research should be conducted to run these types of analyses against different levels and types of de-identification to determine the resulting error levels.

Based on feedback from the operational community we suggest that both a server and a web-ap version should be developed and research on how to keep development in parallel should be done.

Transition Activities

Presentation on de-identification and a basic demo was conducted in Brussels at the ONR sponsored Social Media workshop. Members of NATA and representatives from various programs including IV2 and ICEWS were present.

Presentation on de-identification and a basic demo was conducted at the ONR sponsored demo-day for Dr. Goolsby. Representatives from NSF and STRATCOM were present.

A white paper was sent in to STRATCOM in response to an RFI from STRATCOM.

Milestone Progress Report

The milestones, specified delivery date, and actual delivery date are shown below.

Milestone	Planned Date	Actual Date
1 – Kickoff Meeting	Dec 2013	Jan 2014
2 – Initial Design Specification	17 Jan 2014	17 Jan 2014
3 – Technical and Financial Progress Report I	14 Feb 2014	14 Feb 2014
4 – Technical and Financial Progress Report II	11 Apr 2014	11 Apr 2014
5 – Outbrief on Design Requirements from Operator Community	13 Jun 2014	13 Jun 2014
6 – Final Report	13 Jun 2014	13 Jun 2014

Financial Progress Report

Total Cost under this Firm Fixed-Price Purchase Order: $149, 301.51

Expenses

4 Dec 2013 – 31 Jan 2014	$ 10,893.00
1 Feb 2014 – 31 Mar 2014	$ 26,777.00
1 Apr 2014 – 13 Jun 2014	$ 49,503.00
Total Cost up to 13 Jun 2014:	$ 87,173.00